《心理学报》审稿意见与作者回应

题目：Logistic 加权模型的理论构建与模拟分析
作者：简小珠，戴步云，戴海琦

第一轮

审稿人 1 意见：
《Logistic 加权模型的理论分析与构建》一文提出对传统 Logistic 模型进行改进，增加权重参数以便从难度和重点两个方面进行教育和心理测量。想法独到，有创新性。但有几个问题希望与作者商榷：

意见 1：题目的权重一般是人为确定，也就是专家或老师认定的教学中重点问题。因而，作者认为应该在 IRT 模型中增加这方面的体现。但是，IRT 模型最终是为了估计个体的能力（可能一维或多维），人为设定权重参数意味着人为干预了对个体能力大小的估计，这本身在逻辑上不合理，可能也是为什么国外没有人提出增加此参数的原因。CTT 测量中，对不同题目给予不同的分值，在某种程度上是对题目难度不同的尊重，而 IRT 原有模型已经考虑了题目难度，再加入权重参数，是否还有必要？
回应：本文从两个方面来回答审稿专家的意见：

（一）从分数权重对教育能力培养方向的影响作用

命题教师或专家进行分数赋值、“人为”设定权重参数，不是“武断”的人为设定，也是根据教育规律与教育经验要求来进行的，是客观实践的反映；是一种教育培养方向的引导、评价方向重心转移的体现，而不是简单“干扰”。例如在 20 年前，英语教学中十分强调语法、词汇的掌握，而忽视英语的听力、阅读理解、写作，在英语试卷中语法、词汇的题量大，在整个测验的占比较大，每题的分值为 1 分或 2 分。现在英语教育目标，强调听力理解、阅读理解，因此听力理解、阅读理解的题量较多，分值较大，而相对的，词汇语法题的题量减少，每题的分值为 0.5 分或 1 分，此时，教师、学生都认识到，需要加强听力能力、阅读理解能力的培养训练，而不是花大量时间精力用于学习语法、词汇。

假如现在实践中需要进一步加强提高学生的听力交际水平，而将听力题的 1 分提高到 2 分，并适当增加题量，而将词汇语法的分值由 1 分降低到 0.5 分，那么英语教育将进一步加大英语听力的培养，而在其他方面的培养教学时间将下调；而评价学生英语水平时，也重点强调学生的听力理解能力。

因此，多级记分试题赋予分数权重，不是人为的干预，从教育培养目标、测量目标的角度上看，是一种引导性的作用。试题数量调整、分值调整在学业成就测验的编制、智力测验的编制、修订时是经常遇到的。

即使有个别命题教师主观任意的设定分数权重，偏于教育规律要求，往往会受到其他教育者的“批评”指正。

因此，“人为”设定权重参数，这种干预不是“简单”的干预，而是能引导学生能力的培养发展方向，也有利于测验设计者能实现所需要重点评价的测量方向或能力方向。

（二）试题难度、试题分数权重是试题性质两个不同的维度。

在经典测量理论下，试题难度是计算被试得分比例 P，被试群体能力水平与试题难度相互作用的反映；试题的分数权重作用，也就是在试卷中的贡献作用比例，是由试题满分值 M 及其在

M 占测验总分的比例来反映。

在项目反应理论下的定义，难度 b 是由被试得分比例 P 转换为等距的难度量尺上，同时被试能力也转换到同一量尺上（能力量尺与难度量尺为同一量尺）。而以往的多级记分模型，例如 GRM 将多级记分试题的分数权重也定义到难度量尺上，难度参数与分数权重参数对被试能力估计的作用容易造成混淆。

对于“CTT 测量中，对不同题目给予不同的分值，在某种程度上是对题目难度不同的尊重，而 IRT 原有模型已经考虑了题目难度，再加入权重参数，是否还有必要？”作者理解审稿专家的大意是：难度大的试题，命题教师给予的分值会相对大一些。

作者认为，测验中的一部分试题可以这样操作，如果测验中所有的试题都这样操作，则困难很大。有些试题考察的内容较偏、较难，但不是考察的重点，被试得分率低（难度大）；有些试题内容很重要，大多数学生都会重点学习并掌握，被试得分率高（难度小），以上两种情况，如何来根据难度来赋予试题分值？（或许有研究者会认为，有些内容较偏、较难，在测验中不考察这些。但是，在测验中往往是测验考什么，则教学中就培养什么、学生也学习什么。如果不考那些较偏、较难，教师则不去教这些，学生也不会去学这些。因而完全不考那些较偏的、较难的也是行不通的。）

而且在命题时、在正式测验前，命题教师只能大概预测试题的难度范围。假如命题教师根据预测试题难度小而给予少的分值，预测难度大而给予较大的分值，那么将加大被试得分之间的差距，一些中等能力被试也可能因此得分偏低，不利于测验考察的目的。总之，作者认为测验考察的重点、难点分开操作，有利于实现测量目标。

意见 2：作者在 1.3 节中通过比较了 GRM 和两参数 Logsitic 模型的信息量，就推导出 GRM 模型在分数权重上的不足，逻辑上关系不足。另外，如果想做此类比较，是否还应与 RSM 和 PCM 进行比较。

回应：在项目反应理论下，项目信息量是试题测量属性的最重要指标之一。在本文中论述的含义是，GRM 模型在项目信息量上未能体现倍数关系，因而作为 GRM 模型在分数权重作用上的不足的表现之一，或论据之一，并非绝对的因果推论。作者在论文中也对这部分的有关语句进行了调整修改。

根据以往的研究，在心理与教育测验的能力测验、学业成就测验中，多级记分模型使用情况较多的模型是 GRM。RSM 是 GRM 的延伸发展，PCM 适用的测验情境相对较少。因此，两参数 Logsitic 模型与 RSM 和 PCM 的信息量比较，不是本文所讨论的重点。假如本文中对两参数 Logsitic 模型与 RSM 和 PCM 的比较，可能会造成稿件的篇幅加长，造成内容偏题。

因此，本文只讨论 GRM 和两参数 Logsitic 模型的信息量比较。在本文之外，作者还进一步对 GRM 的信息量和两参数 Logsitic 模型进行了比较，详见“附件 3 GRM 下多级记分题的项目信息量的计算”，得到是同样的结论。

意见 3：作者所提出的权重 Logsitic 模型，设置了一个平均难度表达多级之间的难度，与实际的应用不符。在心理和教育测量中，更多时候，各级别之间难度不同。作者也没有在 2.2 节中给出其模型的完整表达。

回应：难度是被试作答得分的比例的反映。在经典测量理论下，对于多级记分测验试题的难度，也是使用一个得分比例 P 来反映的，即被试群体在该试题上的平均得分除以该试题的满分值的比例。

在以往的多级记分模型中，使用多个难度参数来描述（本质上是被试得分比例转化为难度量尺上的位置参数），这是一种描述难度方式；从理论上讲，这种方式不是唯一方式。

本文中使用平均难度的概念，是试题整体难度的反映，是多级记分试题在难度量尺上的整体

位置参数（也可以看成是经典测量理论下多级记分试题得分比例 P 的反映，在经典测量理论下多级记分试题的难度只有一个得分比例）。而对于被试在各个分数等级之间的差异，本文使用被试得分作为权重来反映。

审稿专家认为：“在心理和教育测量中，（在以往的多级记分模型）更多时候，各级别之间难度不同。”本文作者认为，在以往 GRM 模型下，多级记分试题的各个级别难度之间的间距不同，其本质上是被试群体在各个分数级别上的得分比例不同导致的，而被试群体在各个分数级别上的得分比例受到多个因素的影响，包括试题性质本身、命题教师的分数赋值和评分点设置有关、参与测试的被试群体特征。由 GRM 模型下进行参数估计，可能会发现难度等级的间距不同。然而，这是由被试作答的数据驱动，把分数权重转化到难度量尺上的结果。

但是，换一个角度来看，在试题命制时，多级记分试题的多个等级之间的评分设置，一般是由命题教师（或命题专家）决定的，命题时的基本前提假设为：一道多级记分试题内的多个分值之间是等距的。或者说，不同评分点同样大小的分值，其分值作用是相同的。

对于“作者也没有在 2.2 节中给出其模型的完整表达。”，作者在本文稿件中的“2.3 Logistic 加权模型的项目特征函数推导”部分进行了修改，并较为完整的论述了该模型的函数表达式（用淡蓝色背景显示）。

在原稿中的具体内容如下：

使用被试在该多级记分试题作答的似然函数 $P_{\alpha j}^{\ u_{\alpha j}} Q_{\alpha j}^{\ m_j-u_{\alpha j}}$ 来表示正确作答比例 p，即 $p=k\cdot P_{\alpha j}^{\ u_{\alpha j}} Q_{\alpha j}^{\ m_j-u_{\alpha j}}$，其中 k 可以用二项分布的系数 $C_{m_j}^{u_{\alpha j}}$ 来表示，那么 $p=C_{m_j}^{u_{\alpha j}}\cdot P_{\alpha j}^{\ u_{\alpha j}} Q_{\alpha j}^{\ m_j-u_{\alpha j}}$，这就是本文所要论述的多级记分试题项目特征函数，并将此函数命名为 Logistic 加权模型。其中，$P_{\alpha j}$=1/(1+exp(-1.7$a_j\cdot(\theta_\alpha$-$b_j)$))，a_j 表示区分度参数，b_j 表示平均难度参数。

此时被试在该多级记分试题上恰得 0，1，2，…$u_{\alpha j}$…，m_j 分的正确作答比例分别为：

$$C_{m_j}^{0} Q_{\alpha j}^{\ m_j}，\ C_{m_j}^{1} P_{\alpha j}^{\ 1} Q_{\alpha j}^{\ m_j-1}，\ C_{m_j}^{2} P_{\alpha j}^{\ 2} Q_{\alpha j}^{\ m_j-2}，\ \ldots C_{m_j}^{u_{\alpha j}} P_{\alpha j}^{\ u_{\alpha j}} Q_{\alpha j}^{\ m_j-u_{\alpha j}} \ldots，\ C_{m_j}^{m_j} P_{\alpha j}^{\ m_j} \qquad (2)$$

Logistic 加权模型项目特征函数还有另外一种表达形式，即表示被试在多级记分试题上得 $u_{\alpha j}$ 分和 $u_{\alpha j}$ 分以上的概率表示为

$$P_{u_{\alpha j}}^{\ *}=\sum_{u_{\alpha j}}^{m_j} C_{m_j}^{u_{\alpha j}} P_{\alpha j}^{\ u_{\alpha j}} Q_{\alpha j}^{\ m_j-u_{\alpha j}}，\text{其中}\ 0\le u_{\alpha j}\le m_j \qquad (3)$$

意见 4：第 3.4 节中的参数估计部分没有提出新的内容，只是说用 EM 算法估计，是否有存在的必要。

回应：已删除了该部分的内容，并把该部分的内容移到软件的帮助说明中。

意见 5：文章只是对该模型进行了构想，而没有用程序算法实现，也缺少仿真或实际数据的支持，不符合模型开发类论文的常规写作方法。

回应：已实现了软件编程，实现了新模型的项目参数估计（MMLE/EM 方法），并增加了模拟研究部分。详见修改稿的对应部分“4 Logistic 加权模型的模拟研究”。

审稿人 2 意见

通读全文，整体意见如下（详细意见参见审改稿），仅供参考：

意见 1：作者对 Rasch family models 的理解有所欠缺，进而可能无法准确理解 GRM 与 Rasch family

models 之间在描述多级评分时的区别。

回应：谢谢审稿专家指出的宝贵意见。从不同的视角看，不同的研究者可能将 RSM 会归属于不同的类型。原稿中的对应部分表述不准确，作者已修改。根据 MULTILOG 的关于 RSM 的说明："An extension of Samejima's graded item response model suitable for Likert items is "。

意见 2：由于本文新模型的建构逻辑并不常规，因此需要作者有足够好的写作论述能力和逻辑，以便读者能够理解。但可惜，通读全文，至少本人认为全文的一些关键性问题没有论述的足够清楚，另外有的"假设"的局限性稍强，并且没有在文中给出足够清晰的解释和论证；

回应：非常感谢审稿专家不辞繁琐，耐心、细致的为本文添加了详细的批注，提出了许多宝贵的意见，本文作者收到审稿意见后，认真阅读每一条批注意见，并根据审稿意见进行了修改，并注明了修改说明。

意见 3：根据审稿人的理解，新模型其本质是一个二项式分布：$(a+b)^n=\sum_{r=0}^{n}C_n^r a^r b^{(n-r)}$，作者在文中假设 a=P，b=1-P=Q，则有：$(P+Q)^n=\sum_{r=0}^{n}C_n^r P^r Q^{(n-r)}==1$，进而有 $P(r)=C_n^r P^r Q^{(n-r)}$，即重复 n 次独立试验的成功次数的离散概率分布。其中=号右边的 P 表示单次试验成功的概率，其在 n 次独立试验中是不变的。

之后作者假设 n=m，r=u，进而有 $P(u_j)=C_{m_j}^{u_j}P^{u_j}Q^{(m_j-u_j)}$，则新模型的本质就是 考生答对 mj 个独立的二级评分题的次数的离散概率分布，且答对这些独立的二级评分题目的概率均相等。那么，当给定 θ 时，概率相等则意味着每个题目的难度是相等的。考虑到实际中几乎无法实现编制多道难度完全相等的题目，因此，实际上新模型的本质就转化为： 考生重复作答 mj 次同一道二级评分题目，正确的次数的离散概率分布。

上述内容均为审稿人根据二项式定理推理出的，恰好与作者的模型建构逻辑是一致的："用 m_j 道难度相同的两级记分试题的试题难度来描述或等效于一道满分为 m_j 的多级记分试题的平均难度。"。另外，重复做 mj 次同一题目的信息量当然是做 1 次该题目信息量的 mj 倍。整体上表明审稿人对新模型的理解是比较准确的。

则第 3 个问题是：新模型中暗藏了一个假设，作者并没有在文中提到，比如 m=2、P=0.6，则 P(1)=2*0.6*0.4=0.48，这里 C21=2 表明被试重复 2 次作答同一个二级评分题目，正确作答 1 次的可能有两种：(0,1)和(1,0)，也就是假设考生得 1 分有两种可能路径，当 mj 增加时，中间得分的可能路径更多，作者如何解释该问题？

意见 4：另外，modeling 必须考虑实际应用情境。在现实中，多级评分题目的评分方法通常是设定一系列评分点，比如 multiple-choice item 或 likety type item 的几个选项，或者 constructed-response item 中根据步骤（过程结果）来给分，另外，如果是 constructed-response item 则 modeling 时必须考虑到 rater-effect。不同模型处理的评分方法是不同的，比如 NRM 适合 multiple-choice item 而 GRM 或 PCM 适合 likety type item 和 constructed response item。

根据审稿人对新模型的理解（作者没有考虑或没有说明该问题），新模型应该是针对 constructed-response item 的，或者至少 responses 应该是有序的。那么根据新模型的本质"考生重复作答 mj 次同一道二级评分题目，正确的次数的离散概率分布"，可知，正确次数与得分为 1：1 关系。那么假设当测验编制者在编制一道{0,1,2,3}的题目时，应该如何去确定评分点才能满足新模型的假设（作者回答问题 3 后）？

回应：作者对审稿专家的意见 3 ，意见 4 的核心内容归纳汇总如下：

意见 3：比如 m=2、P=0.6，则 P(1)=2*0.6*0.4=0.48，这里 C21=2 表明被试重复 2 次作答同一个二级评分题目，被试正确作答 1 次的可能有两种：(0,1)和(1,0)，也就是假设考生得 1 分有两种可能路径，当 mj 增加时，中间得分的可能路径更多，作者如何解释该问题？

意见 4：新模型应该是针对 constructed-response item 的，或者至少 responses 应该是有序的。…….可知，正确次数与得分为 1：1 关系。那么假设当测验编制者在编制一道{0,1,2,3}的题目时，应该如何去确定评分点才能满足新模型的假设。

作者具体回应如下：

非常感谢审稿专家对 Logistic 加权模型的理论思想进行深入的探讨分析，帮助了作者对多级记分试题和 Logistic 加权模型进一步的思考与探索。

作者在回答审稿专家的意见 3、意见 4 之前，先论述多级记分试题的评分点结构类型。在心理与教育测验中特别是教育成就测验、智力测验的多级记分试题的评分点类型，可以归纳为以下五种类型。（详见附件 1：多级记分题的评分点结构类型）。

第一类：单纯加权形式的多级记分试题；（Logistic 加权模型可适合）

第二类：并列形式的多级记分试题；（Logistic 加权模型可适合）

第三类：存在一定相依关系的多级记分试题；（Logistic 加权模型可勉强适合，但可能损失测量精度。）

第四类：递进形式的多级记分试题；（Logistic 加权模型可适合）

第五类：以上四种基本类型中的两种或两种以上类型混合而成的类型。

在一份测验中往往包含了以上五种类型的两种或两种以上，如果每一道每一种类型都要细化到具体的评分点上，目前 IRT 下，所有的多级记分模型都是无法同时完全适合这些评分类型。

审稿专家在意见 3 中所描述的评分点情况，可以对应这里的第二类，并列形式的多级记分试题。审稿专家在意见 4 中所描述的评分点情况，可以对应这里的第四类，递进形式的多级记分试题。

例如在一份物理测验中，其中包含了一道多选题、一道解答题。这道道多选题，满分为 4 分，ABCDE 五个选项，答案为 A/B/D，被试选中部分答案（A，或者 AB，或者 AB 等等）给予一半的分值，那么被试得 2 分的可能性作答情况就有很多种，而在同一分测验中的这道解答题中，有两个并行的小题，满分为 6 分，在这两个小题内部都需要分三个严格的步骤进行采分（1，2，3），答对第一步给 1 分，答对前两步给 2 分，全部答对前三步给 3 分。如果要同时适合这两道多级记分题的每一个评分点进行分析，目前 IRT 下的多级记分模型都无法实现适合此测验情境，即使在经典测量理论下也无法区分这些评分点的不同。如果过于纠结于具体的评分点，纠结于在一道多级记分题上同样得 2 分的被试在能力水平上有什么不同，就容易陷入“只见树木，不见森林”的误区。

无论是经典测量理论，还是 IRT，都有这样一个基本假设：在一个测验中（测验中的分测验中），虽然测验中各个试题的测量内容可能是多方面，但总体上的能力维度是一致（或者说能力大致上是单维的），否则，如果各个试题在总体能力维度上不是单维的，那么各道试题的分值之间不能累加（经典测量理论），或者不能在一起进行参数估计（项目反应理论）。

在思维方法上，人们认识世界往往是采取化繁为简，简化概括的方法。在实际教育测验、能力测验中，无论是 CTT 下，还是 IRT 下，在测验命题时、在各道试题上评分操作时都存在以下潜在的基本假设：

（1）在一道多级记分试题上不同评分点上的分值是等距的，或者说贡献作用是相同的。在某一个评分点上的 2 分，与在另外一个评分点上的 2 分，在该道多级记分试题中作用是相同的，否则，该试题的各个评分点之间就不能累加；

（2）由于各个评分点分数作用相同，在某一多级记分试题上各个评分点的得分累加时，在

该道多级记分试题的得分从 0 分到满分最大值，各个分值之间的作用是等距的。

因此，在实际的测验评分过程，在多级记分试题的评分操作中，“忽略”被试具体在某一道试题的哪个评分点上得分，以及具体是哪些评分点的得分累加或组合；而主要关注是在该道多级记分试题的多个评分点的累加得分是多少。

意见 5：作者在文中指出 GRM 存在 4 点不足，那么新模型是否有解决这 4 点不足？

回应：作者在原稿中删除第 1 点， 原稿中的第 1 点是所有 IRT 模型都有可能出现的缺点。GRM 存在的其他 3 点不足归纳如下：

（1）如果在试题的最后一个等级得分的被试太少，往往会严重影响项目难度参数的估计，甚至出现难度参数倒序现象。

（2）当多级记分试题某些中间等级得分上的被试作答信息很少时，需要把这些中间等级进行合并处理，否则 Samejima 等级反应模型就无法对测验数据进行项目参数估计处理分析。

（3）在实际测验中，当试题评分为非连续时，Samejima 等级反应模型下无法处理分析该测验数据

本文通过模拟研究和实测研究发现，在 Logistic 加权模型下，

（a）只设置了一个平均难度参数，不存在着难度倒序问题，也不存在与分数权重作用相混淆的问题。例如“附件 8　实测数据——理数测验被试作答数据 ” 中，最后一道题第 20 题的最大得分 14 分的被试数量很少，在 Logistic 加权模型下，也能估计出项目难度参数。

（b）当试题的中间得分被试作答很少时，Logistic 加权模型可以一并估计，也就是说，可以进行参数估计。例如“附件 8　实测数据——理数测验被试作答数据 ” 中，最后一道题第 20 题的 13 分、12 分上的被试数量非常少，在 Logistic 加权模型下也能估计出项目难度参数。

（c）Logistic 加权模型都可以兼容试题评分为非连续的情况（包括多种非连续性的情况），在本文的模拟研究、和对实测数据的分析，都可以估计非连续性的试题。

修改说明：在原稿中删除第 1 点，并调整修改了有关语句。

意见 6：这点审稿人明白和理解作者欲表达的意思，但不得不说时代在进步，尽管我们不能要求每篇文章的质量都要超过过去，但整体水平和要求肯定是在不断地提升的。因此，审稿人认为“数值计算与计算机编程能力所限”在当今并不是一个足够充分和易被接受的理由，更为恰当的做法或许应该是作者去进一步学习编程或寻找合作者实现参数估计，完成后续必要的模拟研究，这样可从侧面证明参数估计方法可行。

另外，鉴于新模型的建构逻辑并不“通俗”，实证研究（与其他模型的对比）也应该是必要的，这样可说明新模型建构逻辑合理，有现实适用场景。基于一定的研究，才能说明开发新模型的必要性。尽管作者文中提到了，新模型强调了“分数权重”，但作者同样应该意识到以往不强调“分数权重”的模型已经广泛使用了至少 10 年以上且有现成参数估计软件，如果没有充分的理由或较明显的差异，别人为什么要花费功夫学习新模型呢？

回应：针对审稿意见中的几个问题，分别予以回答：

（1）Logistic 加权模型的项目参数估计、测验模拟研究

对于 Logistic 加权模型的项目参数估计（EM 算法），以及模拟研究，目前已解决了项目参数估计、测验模拟算法问题。

作者已按照 MMLE/EM 算法完成了 Logistic 加权模型的项目参数估计，也完成了在 Logistic 加权模型下的测验模拟。为了审稿专家、编辑部、论文读者、其他研究者能够方便的使用 Logistic 加权模型，本文作者提供了可以安装到电脑上的可执行程序（测验模拟软件、项目参数估计软件），并制作了软件的使用说明视频。

这里补充说明一点：测验模拟使用的蒙特卡洛模拟方法模拟测验，不同研究者进行模拟，其

模拟评价指标（ABS、RMSE）波动的范围很小，但不排除偶然（极小概率）的情况出现，模拟评价指标偏差较大。但可以再重新模拟研究过程，再次计算模拟评价指标。

（2）Logistic 加权模型与其他多级记分模型的比较问题

先说明两个模型之间进行模拟比较的一般研究范式：以第一个模型作为基本测验情境进行模拟数据，并在该测验模拟情境进行一些测验情境改变，并导致第一个模型拟合出现问题。与此同时，提出第二个模型，而且第二个模型需要能符合第一个模型及其所对应的测验情境的基本假设（这是模型之间比较的基础），并且能针对测验情境的改变情况能有拟合、纠正作用，从而证明第二个模型能优于第一个模型。最为常见的例子是，两级记分单参数、两参数、三参数模型之间的比较。

Logistic 加权模型的基本前提假设是认为多级记分试题的多级记分是分数加权作用，只有一个平均难度参数；而GRM等其他多级记分模型的基本假设是认为多级记分是分数等级（category），并将多级记分转化到难度量尺上，有多个难度参数。因此，Logistic 加权模型与其他多级记分模型之间没有共同的前提假设，或者说，没有比较的基础。

（3）在实测测验中参数估计情况

本文以数学学业成就测验的一次实测测验数据为例（详见实测数据文件，附件 7、附件 8、附件 9），通过推导，在理想正态分布下 Logistic 加权模型的平均难度参数与 CTT 的难度 P 存在着等式关系 $b=-\varphi^{-1}(p)/\rho$，其中 ρ 是试题得分与总分的二列相关系数，$\varphi^{-1}(p)$ 是多级记分题平均难度 p 的正态密度函数的反函数。

计算测验中的试题得分比例，同时也在 Logistic 加权模型下对测验进行参数估计。当测验中的部分试题被试得分比例 P 为 0.955 时（难度很小），或 0.013 时（难度很大），在 Logistic 加权模型下也能估计出这些的 b 参数，当试题得分比例 P 很大时，则估计出的 b 参数很小；当得分比例 P 很小时，估计出的 b 参数很大，基本上符合。此实例数据分析中，发现 Logistic 加权模型对实测数据估计的平均难度参数与 CTT 下的难度 P 能够大致符合（或很接近）这种等式关系，这也间接的说明 Logistic 加权模型对试题难度参数能够比较准确的反映。

同时，计算该批数据每道试题的点二列相关系数，点二列相关系数是经典测量理论下区分度的一种类型，由结果可知点二列相关系数与 Logistic 加权模型下的 a 参数有高度的相关性，数值上的接近性，这也间接的说明 Logistic 加权模型对试题区分度参数能够准确反映。

表　实测数据中多级记分试题的测量结果

题号	试题满分值	经典测量理论下		Logistic 加权模型	
		点二列相关系数	难度 P	a 参数	b 参数
第 1 题	5	0.314	0.957	0.653	-3.593
第 2 题	5	0.338	0.89	0.428	-3.177
第 3 题	5	0.392	0.918	0.647	-2.816
第 4 题	5	0.437	0.555	0.408	-0.096
第 5 题	5	0.431	0.756	0.446	-1.572
第 6 题	5	0.426	0.787	0.463	-1.805
第 7 题	5	0.3	0.264	0.296	2.54
第 8 题	5	0.297	0.328	0.271	1.984
第 9 题	5	0.309	0.798	0.295	-2.765
第 10 题	5	0.529	0.595	0.598	-0.23
第 11 题	5	0.281	0.32	0.252	2.181
第 12 题	5	0.528	0.574	0.597	-0.108
第 13 题	5	0.212	0.93	0.219	-4.404
第 14 题	5	0.536	0.606	0.618	-0.285
第 15 题	12	0.543	0.82	0.454	-1.533
第 16 题	12	0.504	0.829	0.366	-2.138
第 17 题	14	0.632	0.617	0.563	-0.187
第 18 题	14	0.556	0.496	0.322	0.919
第 19 题	14	0.508	0.171	0.427	2.217
第 20 题	14	0.201	0.013	0.471	3.741

意见 7： 审稿人认为作者要么是混淆了 CTT 和 IRT 中的一些概念，要么是本文没有讲清楚。在 CTT 中，被试的能力=x/M, x 为被试得分，M 为测验总分，即当 M 为定值时，x 与能力成正比。所以，被试答对一个多级评分题后对被试能力的贡献是答对一个 0-1 评分题目对能力贡献的倍数。

但是在 IRT 中，被试得分 x 与能力值 θ 之间仅为正相关，而非固定比例，因此为何答对一个多级评分题目后对被试能力的贡献应该是答对一个 01 评分题目对能力贡献的倍数呢？

如何证明作者的观点或假设呢？

回应：“为何答对一个多级评分题目后对被试能力的贡献应该是答对一个 01 评分题目对能力贡献的倍数呢？”

这是作者的理论上的推论，其推断来源于以下两个方面：

一是，在测验命题时、在各道试题上评分操作时都存在以下潜在的基本假设：（1）在一道多级记分试题上不同评分点上的分值是等距的，或者说贡献作用是相同的。（2）由于各个评分点分数作用相同，在某一多级记分试题上各个评分点的得分累加时，在该道多级记分试题的得分从 0 分到满分最大值，各个分值之间的作用是等距的。

二是，Logistic 加权模型下的项目信息量：

$$I_j(\theta)=\frac{m_j\cdot[P_j'(\theta)]^2}{P_j(\theta)\cdot Q_j(\theta)}=m_j\cdot\frac{[P_j'(\theta)]^2}{P_j(\theta)\cdot Q_j(\theta)}$$

从数学函数上看，该项目信息量公式体现了一道多级记分试题多级记分试题的项目信息量是一道难度值相当的两级记分试题的 m_j 倍，也就是说，Logistic 加权模型下多级记分试题能提供较多的项目信息量，从而有较高的测量精度。而在 GRM 下，一道多级记分试题提供的项目信息量

仅只比一道难度相近的两级记分试题所提供的项目信息量多一些，在项目信息量上并不能体现多级记分的权重倍数。

修改说明：对原稿中语句修改，修改为“第一，从项目信息量的角度来看，项目信息量是试题属性最重要的测量指标之一。在 GRM 模型下，一道多级记分试题提供的项目信息量仅仅比一道难度相当的两级记分试题的项目信息量仅多一些，也就是说，多级记分试题的项目信息量并不是两级记分试题的倍数关系。”

第二轮

审稿人 1 意见

意见 1：作者对论文做出了较大修改，特别是增加了仿真和实际数据的支持，使文章的质量得到提升。对于是否有必要增加“权重参数”，作者给出了更多的解释和回应。

回应：审稿专家的意见“对于是否有必要增加‘权重参数’”，作者的观点是：在测验中试题分数权重参数（试题满分值）在测验命题时就已经确定了，已客观存在，而不是审稿专家提出的 “是否有必要增加‘权重参数’”。以下具体解释。

在 CTT 下，例如在一份能力测验（或学业成就测验）中，题型一的试题满分值为 1 分，题型二的试题满分值为 3 分，题型三的试题满分值为 5 分，在计算被试总分时累加各个试题的得分，此时试题满分值产生了分数算术加权作用，此时这三种题型上的试题满分值本质上是一种分数加权（算术加权，试题满分值也就是分数权重参数），这是一般共识。在学业成就测验中，在教学内容中分为教学重点、难点，相对应的，在测验考查中也分测验重点、难点。在命题时，命题教师根据教学内容要求、考试大纲要求、教学规律经验而设置试题满分值（分数权重）。因此，试题满分值（分数权重作用），与试题难度属性一样，是试题的两个不同测量现象，是客观存在的。

同样的，IRT 数学模型对应的也有难度参数、分数权重参数。在 GRM 模型中，试题满分值是借助多个递增形式难度参数表达（难度等级形式，分数等级形式）；而本文提出的 Logsitic 加权模型使用几何加权形式（指数形式）来表达，因为在 IRT 下似然函数运算公式是使用函数连乘的方式，因而分数加权作用就需要使用几何加权形式（指数形式）。因此，Logsitic 加权模型并不是“额外”引入了权重参数。正如前一段所说，试题的分数权重参数（试题满分值）在试题命题时就确定了，即已经存在的，只不过，试题满分值（试题分数权重参数）在 GRM 模型、Logsitic 加权模型这两个数学模型中有着两种不同表达形式，一个为分数等级形式（难度等级形式），一个为分数几何加权形式。

对于审稿专家意见：“作者给出了更多的解释和回应。”作者在第一轮回复中主要是回答了测验中为什么需要进行分数加权（即测验中为什么需要分数权重参数）。

意见 2：并用英语考核“词汇”还是“听力”为例说明权重的重要性，然而“词汇”和“听力”并非一种能力，没用理由用一个 IRT 模型进行计分。

回应：审稿专家的这一意见表面上看是对的，然而，实际上不完全如此。在实际考试中，我们将听力部分得分，与其他词汇、语法、阅读部分的得分累加成一个总分（英语成绩得分），那么就意味着听力与其他部分、以及总分（英语基本能力、或英语综合能力）都是看作在同一个能力维度上，那么在一个能力维度上就可以使用一个 IRT 模型来计分。否则，如果测验听力，与各个部分不是作为一个能力维度，那么在经典测量理论下，各个部分的得分也不能相加为一个测验总分。因此，在第一轮回复中在一个测验中听力、词汇、语法等部分使用一个 IRT 模型计分，该例子基本上还是可行的。

当然，如果将第一轮回复中的英语测验的听力、词汇、语法、阅读理解等，修改为某一学科测验中的内容模块 A、内容模块 B、内容模块 C、内容模块 D 等，那么，审稿专家和其他研究者

可能更易于理解接受。

意见 3： 还有几个写作上的建议：一是将第 4 节“Logistic 加权模型的模拟研究”与第 3 节内容对调，才符合此类论文撰写的一般规则；二是 3.1 节和 3.2 节标题中均出现了问号“？”，这也不符合文章撰写规则；三是结论部分应为文章点睛之笔，但目前文字虽然较多，却没有体现出研究的亮点。

回应：非常感谢审稿专家的宝贵意见。已按照审稿专家的意见，在原文中进行了相应的修改。

审稿人 2 意见

意见 1：作者较好地回答了审稿人的问题，尽管文章的主要思想并非主流，但我们应允许新思维的存在和出现。

建议修改稿中大幅增加实证研究的内容，模拟研究主要在于探究参数估计方法的可行性，而实证研究才是展现模型用武之地的地方。

回应：谢谢审稿专家的肯定意见，将按照审稿专家的意见，在原文适当增加了一些分析论述。

作者认为，本文已使用 Logistic 加权模型实现了对实测数据的试题参数估计，也就是说，实测数据的项目参数估计，作为最为关键的应用问题都解决了。因此，其他测验应用问题，包括 Logistic 加权模型下的测验等值、被试能力估计等实证应用问题就比较容易解决。

如果本文中再增加一个实测应用研究，那么可能需要增加实测研究的研究设计、研究方法、研究结果、分析与讨论等大篇幅内容，很容易就变成了论文中包含了“第二篇研究论文”，导致此论文篇幅过长。

审稿人 3 意见

第三位审稿专家在原稿件中给出了 26 条批注，一部分批注在此第二轮回复意见中进行回复；另外一部分批注较为简要，则直接在原稿批注中直接回复。

意见 1：（对应批注 1）实测测验数据中的“最后一道题第 20 题的最大得分 14 分的被试数量很少，在 Logistic 加权模型下，也能估计出项目难度参数”。（以上为原稿第一轮回复中的内容）

GRM 也能估计，只是像作者所说，其参数估计偏差较大。而新模型在该情况下虽能估计出难度参数，但其偏差是否也是和 GRM 一样大呢？作者没有进行验证，如果是的，那么这不能算是解决了 GRM 的缺点。请作者解释。

回应：（1）可能审稿专家没有注意到，如果仔细查看这批实测数据（第一轮回复的附件 8、或附件 9），就会发现这一批数据的前面一共有 14 题为客观题得分形式，这些试题的评分只有 0，5 两种得分情况，从理论上看，这种得分出现跨越情况与其他试题的得分情况一起，使得整个测验得分在 GRM 下是无法估计的。以上仅是作者的个人观点，对于这批数据在 GRM 下是否能进行估计、如何进行估计，还希望能进一步交流探讨。

（2）即使假设这批数据在 GRM 下能进行估计，在 GRM 下第 20 题能估计 14 个难度参数（难度参数依序增大），然而，也难以找到在 GRM 下计算偏差的依据（或参照标准）。只有在模拟测验中，模拟测验数据有难度初始值、难度参数估计值，此时可以计算偏差。

在 Logistic 加权模型下，第 20 题的难度参数估计值也是一个实测数据的参数估计值，也同样是难以找到计算偏差的依据（或参照标准）。

意见 2：（对应批注 9）如果作者也能推导出 Logistic 加权模型的难度参数与 CTT 参数之间的关系，将会进一步丰富新模型的理论构建。

回应：在二级计分模型下，Lord（1968）在理想正态分布下推导了两参数模型与 CTT 的难度参数

的等式关系（可参见：Lord F. M. (1992). 心理测验分数的统计理论（叶佩华 译). 福建教育出版社. 第 427-430 页 ）。同样的，在理想正态分布下可推导多级记分题下 Logistic 加权模型的与 CTT 的难度参数等式关系，其推导过程较为复杂，最后的等式关系如下（其他研究者也可以具体验证）：

Logistic 加权模型与 CTT 的难度参数的等式关系为：$b_j = -\frac{\varphi^{-1}(p)}{\rho_j}$

区分度参数的等式关系为：$a_j = \frac{\rho_j}{\sqrt{1-\rho_j^2}\sqrt{m_j}}$，

其中，m_j 为试题满分值，$m_j \geq 1$；

a_j，b_j 分别为 IRT 下的多级记分题的区分度、平均难度参数；

$\varphi^{-1}(p)$ 为多级记分题平均难度 p （CTT 下多级记分题的难度参数 $p = \bar{x}/m_j$ ）的正态密度函数的反函数，或者说，多级记分题的平均得分率 p 转换标准正态分布下的 Z 分数；

ρ_j 是项目得分与试卷总分的二列相关系数（CTT 下区分度的一种计算方法）；

意见 3：（对应批注 11 ）在心理与教育测验中多级记分试题的试题属性主要表现为：（1）试题知识考查重要性程度的加权；（2）试题的平均难度。（原稿内容）
有参考文献吗？还是作者自己总结的。请给出这样分类的说明。
这点很重要，因为这是作者新模型的立意所在，作者在后面多次强调现有多级模型无法体现分数加权作用，那么我们为什么要考虑呢？理由是什么？对于作者在后面给出的解释，我的疑问还不少，详见 1.2 部分接下来的几个问题。
回应：原文中的这句话是作者自己总结的，是作者对测验中多级记分试题的测量目标、测量属性的概括。在这第二轮修改中，作者将这句话修改如下：

在心理与教育测验中多级记分试题的试题属性主要表现为：（1）用试题满分值来表达试题知识考查重要性程度的加权作用；（2）用试题的平均难度来表达被试群体在多级记分试题上的得分比例。

（1）对于多级记分题的第一点属性，作者解释如下：多级记分题（即测验中的“大题”，特别是测验后面的大题）一般都是测验考查的重点内容、或较为重要内容，这些大题的分数往往相对较多，在一定程度上反映了考查内容重要性的分数权重。如果不是很重要的知识点，那么在测验中则不能安排大题（较大的试题满分值）来考查，否则，我们会认为命题教师在编制命题时就出现了失误偏差。

而且，在一份测验中，教学内容、考查内容的重点，既可以通过增大某一道试题的满分值（权重参数）体现，也可以通过增加在某一内容模块的试题题量来体现，该内容模块的所有试题的满分值累计可以得到模块分数（或模块总分）；那么某一模块内容的重要性最终通过模块总分在整份测验总分的占比（比重）来体现。

在原稿中，也说明了“教学内容具有重点（重要性）、难点（难易程度）这两个属性，相对应的，试题也具有考查重要性、难易程度这两个属性。”一般来说，教学内容越重要，那么考查该内容的模块总分比例就越大。模块总分比例的增大，可以通过增大试试题满分值来体现，也可以通过增加考查该内容模块的试题题量来体现。

由以上，本文概括多级记分试题的第一个试题属性：用试题满分值来表达试题考查重要性程度的加权作用。

（2）对于多级记分试题的第二个试题属性，也是作者自己归纳的。并在原稿中补充了以下内容：“由于多级记分题的评分点结构或评分等级相对复杂，但人们为了简化思维认识，可以使用平均难度表达被试群体在多级记分试题上的得分比例。”

审稿专家的意见：“后面多次强调现有多级模型无法体现分数加权作用，那么我们为什么要考虑呢？理由是什么？”本文在“审稿意见 5（对应批注 15）”中对原文例子的拓展论述，也就是GRM 模型无法体现分数加权作用的论证。这部分内容较多，具体内容详见“审稿意见 5（对应批注 15）”的回复内容。

意见 4：（对应批注 12 、批注 14）其中批注 12 为：“多级记分试题的分数权重就是体现对所考查知识点的重要性加权”（原稿内容）不太理解。作者想表达的是分数越高的题目所考查的知识点越重要吗？本人认为有些二级计分题目的重要性不见得比多级题目低，只是某些知识点适合二级计分而已。

批注 14：试题知识考查重要性则需要通过分数权重来体现，即需要另一个参数来体现，本文中称为权重参数。（原稿内容）

问题 1：请作者再仔细斟酌。本人认为知识点的重要性不能简单的从分数的多少来判断，正如作者前面定义的：“重点是指所考查知识能力（或心理因素）的重要性，是由知识能力（或心理因素）的基础性、重要性、应用广泛性来决定的。”一道二级计分题目同样可以考查重点，不见得非得用多级题目才能考查。

问题 2：由上一问引出该疑问。请作者解释“权重参数”的含义。因为原来我们熟知的各种 IRT 模型，模型参数的含义是十分明确的，我在这里不能很好地理解作者提出的“权重参数”的含义。另外，如果二级计分的重要性比多级计分还要高，这个权重参数又如何理解？

回应：在一份测验中，测验考查内容的重点，可以通过增大每一道试题的满分值（分数权重参数）体现，也可以通过增加测验试题的题量来体现，并最终通过某一内容模块的模块总分在整份测验总分的占比（比重）来体现。

在许多情况下，测验中的重要内容通过大题来考查，不重要的内容通过小题来考查，但是这种情况是相对的。例如一份测验分为多个模块：模块 A、模块 B、模块 C、模块 D……。对于某一重要内容模块——模块 A，选择题安排了 3 道试题（2 分一题），填空题安排了 1 道试题（4 分一题），应用题安排了 1 道试题（12 分一题）；而另一个相对不重要的内容模块——模块 C，则只安排了一道填空题小题（4 分），内容模块 C 在该测验中没有再安排其他试题。 此时，模块 C 的一道填空题 4 分，多于模块 A 的一道选择题小题 2 分，此时，我们不能说模块 A 由于考查了一道单选题而重要性就下降了，重要性不如模块 C。

因此，审稿专家提出的测验现象（包括批注 12、14 中的“二级计分题目的重要性不见得比多级题目低”，“如果二级计分的重要性比多级计分还要高”、“一道二级计分题目同样可以考察重点”）是存在的。但是，如果一道二级计分题目的重要性很高的话，那么命题教师在该二级计分题目所对应的模块，可能增加该模块的题量（即增加二级计分题量，或者增加多级记分题量），使得该模块的模块总分在整份测验总分的比重较高。否则，如果重要内容模块却只考了一道小题（只考了 1 分或很少的分值），那么命题教师就存在着命题上的偏差，考查重点失偏，此份测验就无法达到测验目标。

在一般测验情境下，试题的满分值作为试题的分数权重参数，在测验总分中起到了分数权重作用，通过模块总分（该模块所有试题的满分值累加和）来反映知识点或内容模块重要性，这是命题、测验编制时的基本命题思想。

因此，在 CTT 下，试题满分值作为试题的分数权重参数（算术加权形式），被试在各个试题上的得分通过算术累加得到测验总分，其测量意义还是比较明显的。如果在 IRT 下，被试能力估计的似然函数运算公式是使用函数连乘的方式，那么试题的权重参数形式使用几何加权形式。

修改说明：在原文中增加以下内容：测验考查知识点（内容模块）的重要性，可以通过增大每一道试题的满分值（分数权重参数）体现，也可以通过增加测验试题的题量来体现，并最终通过某一内容模块的模块总分在整份测验总分的占比（比重）来体现。

意见 5：（对应批注 15）“在 IRT 中，潜质与题目难度共同同一量尺且相对独立是一个基本假设。在同一能力量尺上，可认为 A 和 B 均跨过了 b=2.5 这个点，而 C 没有跨过。因此，对 C 的能力估计值小于 A 和 B 是符合逻辑的。其实 C 的作答结果(8)就相当于 A 中的 0，只不过，C 答的这道题把 A 中的 0 这一类又进行了更为细致的划分。所以提供的信息量可能会有所增加。

该例子，并未让审稿人直接明白作者想强调的东西，而是，“我猜”作者是想说，为什么 C 答对了 8/9 的题目，其能力估计值还不如答对一个 1/1 的题目后的 A 的估计值高。但正如审稿人上面所述，C 比 A 能力估计值低是符合逻辑的。或者，至少作者的描述并未让审稿人对“C 比 A 能力估计值低”这句话的逻辑产生质疑。”（第一轮审稿专家的审稿意见）

首先作者没有回答上一位审稿人对于该举例的疑问。其次，我也感觉这个例子有问题。9 分题目的最高等级难度和 2 分题目的难度是一样的，那么被试答对任何一题，能力估计值必然一样。但被试如果在 9 分题目上得 8 分（其对应的难度会小于 2.5），那么得到一个低一点的能力估计值也是必然的。这并不是由作者所谓的分数权重造成的。

再次，基于作者构建的 Logistic 加权模型，我举个例子：第一题是 b=3 的二级题目，第二题是 b=2 的多级题目。前者难度高，假设很重要，但它的分数权重小；后者难度低，重要程度假设不如第一题，但它的分数权重大。两个相同能力的被试甲和乙，甲做对第一题做错第二题，乙相反。这时在运用 Logistic 加权模型更新两人的能力时，难度和分数权重的作用方向是相反的。这个时候怎么处理？

回应：（一）对于原文的该例子，两位审稿专家都提出了不同程度的疑问。这里将原稿中的例子放到整个测验情境中进行说明。

假设仍然在 GRM 模型下，两个被试（甲、乙）在一份英语试卷上的作答得分大部分相同，只有一道词汇题、一道作文题的得分不一样，词汇题满分为 1 分，难度为 2.5；作文题的满分为 10 分，10 个难度依次为-2.0，-1.5……2.0，2.5，最高难度为 2.5。假设被试甲词汇题得 1 分，作文题为 0 分（被试甲作文很差，或未作答），被试乙词汇题得 0 分，而作文题得分为满分 10 分，那么在 GRM 模型下被试甲和被试乙的能力估计值是一样的（此时未能有效的区分被试能力）。间接的说，一道作文题的作用（分数权重很大），与一道词汇题的作用是一样的。此时，在 GRM 模型下估计就没有达到测验命题时的测量目标。因为作文能力是测验中最重要的能力，在命题时大幅增加作文题的满分值，是为了增大该作文题对被试的区分能力，让作文能力强的学生得到更高的分数。

接下来，再进行以下两种测验情境假设：

第一种假设：假设该例子的测验情境基本都不变，所有被试在测验上所有试题的作答情况（被试作答内容）都不变，包括被试甲、乙在词汇题、作文题上的作答内容也不变。唯一变化的是，命题教师直接增加作文题的满分值为 20 分（相对应的，所有被试在作文题上得分都乘以 2，即原来被试得 3 分，则现在为得 6 分，作文得分等级仍为 10 个等级）。命题教师增加作文分是为了进一步强调作文能力方面的重要性，让作文能力强的学生得到高分，测验目标是所有学生重视和加强作文能力的培养。

由于所有被试在该测验的作答内容不变，那么被试在各个试题上的得分比例 P 不变，包括作文题上的各个评分等级上的得分比例 P 也不变，因而可以进一步推论，该测验在 GRM 下的难度参数都不变，该作文题在 GRM 下的最高难度参数依然为 2.5，而词汇题的难度参数也是 2.5，此时在 GRM 下被试甲、被试乙的能力估计值依然还是一样。也就是说，此时在 GRM 下的能力估计达不到命题教师的命题目标，无法让作文能力强的学生得到更高的分数。如果将作文满分值进

一步增加到 40 分，60 分时，并且假设所有被试在该测验的所有作答内容不变，而在 GRM 下被试甲、被试乙的能力估计值依然还是一样。

第二种假设：让该作文题的满分值的难度参数进一步加大，比该词汇题的难度参数大许多。这里就假设作文题满分为 18 分时，在 GRM 模型下最大难度参数可能是 7.5 或更大值，那么对被试能力估计值会有怎么样的影响呢？以下再分两方面来论述：

一、增大 GRM 模型的最高难度参数的对能力估计值的影响很小

假设一个中等能力被试（能力值为 0.0）在一个测验上作答结束，并估计出一个能力值θ_0，在该测验结束后，再给被试额外增加一道多级记分试题的测试，为了考查被试作答不同难度的试题，让该多级记分题的难度参数在[-7.2, 7.2]之间变化。被试作答该多级记分题后，其能力估计值为θ，由θ再减去参照值θ_0，得到$\theta-\theta_0=\Delta\theta$，即能力变化幅度。

根据 Bock 和 Aitkin（1981）提出项目反应理论的基本假设：项目与项目之间相互独立，项目与被试作答相互独立，被试作答与被试作答之间相互独立。因此可以认为，θ与θ_0相减所得的能力变化幅度，是由被试作答该多级记分题后产生的。

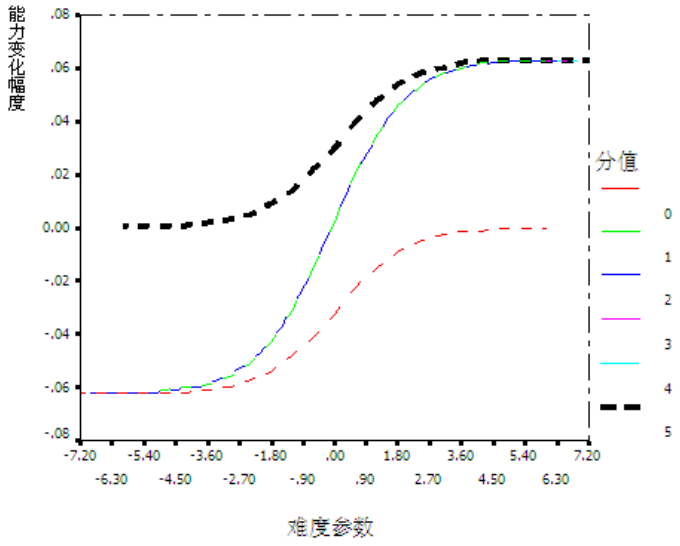

图1　GRM下能力估计值的变化幅度（在MULTILOG软件下估计）

当某一被试在额外答对一道多级记分题并得满分 5 分，当多级记分试题的最高难度参数为 2.7 时，其能力变化幅度为 0.058；当该多级记分题的最高难度参数为 7.2 时，其能力变化幅度为 0.064，几乎变化不大。特别是在难度参数从 3.60 增大至 7.20 之间时，被试能力的变化幅度几乎没有改变，也就是说，GRM 模型下，当多级记分题的最高难度参数从 2.7 变化至 7.2 时，最高难度参数的增大对被试能力估计值的影响（能力变化幅度）很小。

二、增大 GRM 模型的满分值或最高难度参数对项目信息量的影响很小

在 GRM 下，增大多级记分题的满分值，或增大 GRM 的最高难度，计算其项目信息量，详见第一轮审稿回复的附件 3《附件 3 GRM 下多级记分题的项目信息量的计算》。在附件 3 中的主要结果，如下图。

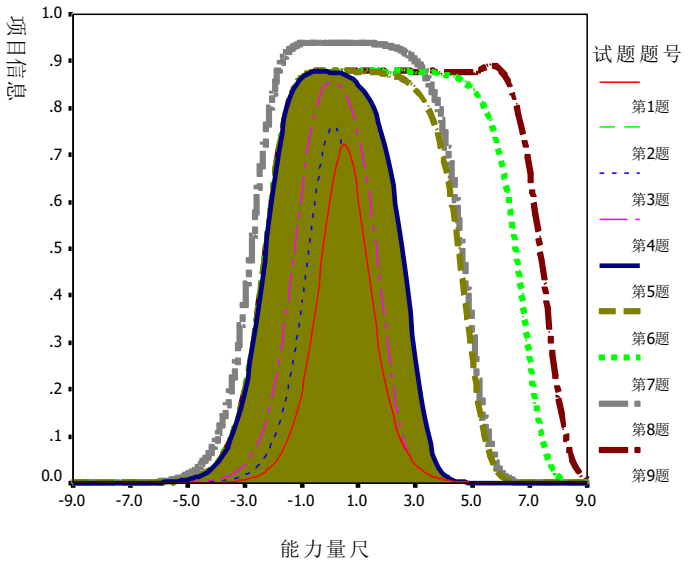

图 2 九道试题在 GRM 下的项目信息量

在 GRM 下计算项目信息量，以第 4 题（满分值为 5 分）的信息量曲线为基本参照曲线。

对于图中的第 9 题，试题满分值增大为 18 分，并加大试题的最高难度参数值为 8.2，该多级记分试题的难度参数依次为:，-2.5，-1.5，-0.5，0.5，1.5，2.5，3.5，4.5，5.5，5.8，6.1，6.4， 6.7，7.0，7.3，7.6，7.9，8.2。但是，多级记分题的项目信息量在整个能力量尺上的信息量最大值仅仅产生了非常小的增幅。

对于图中的第 8 题，增大试题满分值为 17 分，增加试题难度参数个数密度（即每个难度参数值之间间距很小），而且试题的最高难度参数值为 5.5，该多级记分试题的难度参数依次为：-2.5，-2.0，-1.5，-1.0，-0.5， 0.0，0.5，1.0，1.5，2.0，2.5，3.0，3.5，4.0，4.5，5.0，5.5。但是，第 8 题的项目信息量在被试能力量尺上的最大值，比第 4 题的最大项目信息曲线仅仅增大了 0.1 左右（在被试能力范围[-2, +3]）。

总之，由以上图 2 的九道试题的项目信息量情况，可知，在 GRM 模型下，成倍的增大多级记分题的满分值，或成倍的增大最高难度参数，对项目信息量的影响都很小，那就意味着对测量精度影响很小。

意见 5(续)：基于作者构建的 Logistic 加权模型，我举个例子：第一题是 b=3 的二级题目，第二题是 b=2 的多级题目。前者难度高，假设很重要，但它的分数权重小；后者难度低，重要程度假设不如第一题，但它的分数权重大。两个相同能力的被试甲和乙，甲做对第一题做错第二题，乙相反。这时在运用 Logistic 加权模型更新两人的能力时，难度和分数权重的作用方向是相反的。这个时候怎么处理？

回应：第一部分：试题内容重要性与试题分数权重大小的关系分析

前面已经论述了，在一份测验中，测验考查内容的重点，可以通过增大每一道试题的满分值（分数权重参数）体现，也可以通过增加测验试题的题量来体现，并最终通过某一内容模块的模块总分在整份测验总分的占比（比重）来体现。

对于审稿专家中的举例，第一题 b=3 的二级题目，“前者难度高，假设很重要，但它的分数权重小；”作者认为，如果第一题该二级题目的内容很重要，那么在命题时就应该（增加该内容模块对应的二级记分题的题量，或者增加该模块对应的多级记分题的题量），从而该内容模块的模块总分在测验总分中占相对很大的比例。按正常的测验命题情况，那么第一题很重要的话，那么第一题所在的内容模块应该还对应了其他题量的试题，使得该内容模块的总分比例比较大。

第二题是 b=2 的多级题目，如果该内容模块不重要，则可能多级记分分值相对也不大，分值

比第一题二级计分题目所在内容模块的模块总分要低。假如第二题所在的内容模块不重要，但该内容模块在这个测验总分的比例很大，那就是命题教师出现命题失误了。

因此，从命题角度来看，试题内容重要性、与试题分数权重（指模块总分权重）应该需要在测验总体上表现为一致。

第二部分：试题难度与试题分数权重（试题内容重要性）的关系分析

对于审稿专家提到的“难度和分数权重的作用方向是相反的”。作者对试题难度、试题分数权重的关系可能出现类型，分为四类典型的组合：

A、试题难度大、分数权重大，这往往是测验中重点考查的大题、重要性的大题，往往放置测验中的中部位置，或最后压轴位置；

B、试题难度大、分数权重小，例如试卷中的部分填空题、选择题，试题满分值小，这样设置的理由或原因是：（1）有些知识点较偏，被试答不出来；（2）或者有些内容不重要、且难以理解掌握，大部分被试作答错误；（3）一大部分考生理解失误或其他原因，导致答错或答不出来，等等。这些原因导致了得分比例小，试题难度大。

C、试题难度小、分数权重大，这在试卷中一些考查重要知识点的试题可能出现，由于一些重要知识点，教师重点讲解，学生也重点练习与掌握，此时大部分被试都掌握了，那么试题得分比例高，试题难度小。

D、试题难度小、分数权重小，这种情况包括：（1）测验中的一部分基础题中常出现；（2）在考查不重要知识点时也常出现；（3）某一重要知识内容模块考查了多道试题（包括多级记分题、两级记分题），在这些试题中可能出现一些试题难度小、分数权重小的现象。

这里对审稿专家的举例进一步进行分析，第一题是 b=3 的二级题目，满分值为 1，第二题是 b=2 的多级题目，满分值为 2。 如果被试甲、乙都是同样能力，还可以细分以下两种情况：

（1）被试甲、乙同为低能力被试（或中等能力被试），那么被试甲、乙很可能（非常大的概率）同时答错这两道试题（b=3、b=2 高难度试题）。

（2）被试甲、乙都是同为高能力被试时，出现审稿专家所说的这种情况“甲做对第一题做错第二题，乙相反。”的概率非常大。此时这两道试题对被试能力值的作用，需要看这两道题对这两个被试提供的项目信息量大小来进行分析。

而第一题如何体现所在内容模块的重要性，还需要依赖于该内容模块中的其他试题一起共同作用。

意见 6：（对应批注 18）用什么方法表示不是重点，但都要说明这样表示的道理和参数的含义。二级计分 IRT 模型中的难度参数 b 的含义是正确作答概率恰为 0.5 时所对应的能力值是题目的难度。GRM 的难度等级也能够刻画能力与难度之间的关系，例如，难度等级依次为-2，-1，0，1，2 五个等级，若被试的能力为-0.5，那么他的能力就对应第二等级难度，若能力为 1.3，能力就对应第四等级难度，这些模型中的难度参数都是有具体含义的。

请作者解释新模型中“平均难度”这个参数的具体含义，而不是简单的说一下，我认为这样是不行的。

回应：感谢审稿专家的宝贵意见。作者在原稿的后续部分“3.2　对 Logistic 加权模型下只用一个难度参数的讨论”也论述了使用一个平均难度参数的含义。根据审稿专家的意见，在此修改稿中将进一步论述。

在经典测量理论中，一道多级记分试题包含多个小题或多个评分点，计算此道试题难度(或得分率)使用公式 $P=(\sum_{i=1}^{N}x)/(N\cdot m_j)$，此得分率 P 是平均难度的涵义。

在 IRT 理论下，多级记分题的平均难度参数的参数含义与二级计分试题的难度参数含义相似，表达的当被试在某一多级记分题上的得分为中间得分及以下的期望概率累加和为 0.5 时（相对应

的，得分为中间得分及以上的期望概率累加和也是为 0.5），此时该被试的能力估计值即是该多级记分题的平均难度参数。换句话说，当被试能力估计值与某一多级记分题的平均难度参数大小相等时，该被试在该多级记分题上得分为中间得分及以下的期望概率累加和为 0.5。作者将在原稿中增加以上内容的表述。

具体推导过程如下：

被试在该多级记分试题上恰得 0，1，2，…$u_{\alpha j}$…，m_j 分的正确作答比例分别为：

$C_{m_j}^0 Q_{\alpha j}{}^{m_j}$，$C_{m_j}^1 P_{\alpha j}{}^1 Q_{\alpha j}{}^{m_j-1}$，$C_{m_j}^2 P_{\alpha j}{}^2 Q_{\alpha j}{}^{m_j-2}$，…$C_{m_j}^{u_{\alpha j}} P_{\alpha j}{}^{u_{\alpha j}} Q_{\alpha j}{}^{m_j-u_{\alpha j}}$…，$C_{m_j}^{m_j} P_{\alpha j}{}^{m_j}$

（一）当多级记分题的满分值为奇数（以 5 分为例），假设某一被试能力估计值为$\theta = 0.3$在该多级记分试题（平均难度参数$b = 0.3$）上得到中间得分及以下（即 2.5 分及以下得分）时的期望概率，具体计算如下：

该被试在该多级记分题上恰得 0，1，2，3，4，5 分的期望作答概率（有 6 种情况）分别为：

$C_5^0 Q_{\alpha j}{}^5$，$C_5^1 P_{\alpha j}{}^1 Q_{\alpha j}{}^{5-1}$，$C_5^2 P_{\alpha j}{}^2 Q_{\alpha j}{}^{5-2}$，$C_5^3 P_{\alpha j}{}^3 Q_{\alpha j}{}^{5-3}$，$C_5^4 P_{\alpha j}{}^4 Q_{\alpha j}{}^{5-4}$，$C_5^5 P_{\alpha j}{}^5$

其中$P_{\alpha j} = 0.5$，中间得分为 2.5（在 2 分、3 分这两个位置的中间）时，此时被试的累加期望概率是包括前三项，$C_5^0 Q_{\alpha j}{}^5$，$C_5^1 P_{\alpha j}{}^1 Q_{\alpha j}{}^{5-1}$，$C_5^2 P_{\alpha j}{}^2 Q_{\alpha j}{}^{5-2}$，即

$$
\begin{aligned}
&C_5^0 \bullet (1-0.5)^5 + C_5^1 \bullet 0.5^1 \bullet (1-0.5)^{5-1} + C_5^2 \bullet 0.5^2 \bullet (1-0.5)^{5-2} \\
&= 1 \bullet 0.5^5 + 5 \bullet 0.5^5 + 10 \bullet 0.5^5 \\
&= 0.5
\end{aligned}
$$

特例情况：当多级记分题的满分值为 1 分时，假设某一被试能力估计值为$\theta = 0.3$在该多级记分试题（平均难度参数$b = 0.3$）上得到中间得分及以下（即 0.5 分及以下得分）时的期望概率，具体计算如下：

该被试在该多级记分题上恰得 0，1 分的期望作答概率（有 2 种情况）分别为：

$C_1^0 Q_{\alpha j}{}^1$，　$C_1^1 P_{\alpha j}{}^1$

其中$P_{\alpha j} = 0.5$，中间得分为 0.5（在 0 分、1 分这两个位置的中间）时，此时被试的累加期望概率是包括前 1 项，即

$$
\begin{aligned}
&C_1^0 \bullet (1-0.5)^1 \\
&= 1 \bullet 0.5^1 \\
&= 0.5
\end{aligned}
$$

（二）当多级记分题的满分值为偶数（以 6 分为例），假设某一被试能力估计值为$\theta = 0.3$在该多级记分试题（平均难度参数为$b = 0.3$）上得到中间得分（3 分）是的期望概率，可以计算出来为 0.5。具体计算如下：

该被试在该多级记分题上恰得 0，1，2，3，4，5，6 分的期望作答概率（有 7 种情况）分别为：

$C_6^0 Q_{\alpha j}{}^6$，$C_6^1 P_{\alpha j}{}^1 Q_{\alpha j}{}^{6-1}$，$C_6^2 P_{\alpha j}{}^2 Q_{\alpha j}{}^{6-2}$，$C_6^3 P_{\alpha j}{}^3 Q_{\alpha j}{}^{6-3}$，$C_6^4 P_{\alpha j}{}^4 Q_{\alpha j}{}^{6-4}$，$C_6^5 P_{\alpha j}{}^5 Q_{\alpha j}{}^{6-5}$，$C_6^6 P_{\alpha j}{}^6$

被试得分为中间得分（在 3 分这个位置的中间位置）时，此时累加期望概率是包括该数列 7 项中的前 3.5 项，即包括了前三项、和$C_6^3 P_{\alpha j}{}^3 Q_{\alpha j}{}^{6-3}$的一半，由此可得：

$$C_6^0 \bullet (1-0.5)^6 + C_6^1 \bullet 0.5^1 \bullet (1-0.5)^{6-1} + C_6^2 \bullet 0.5^2 \bullet (1-0.5)^{6-2} + 0.5 \bullet C_6^3 \bullet 0.5^3 \bullet (1-0.5)^{6-3}$$
$$= 1 \bullet 0.5^6 + 6 \bullet 0.5^6 + 15 \bullet 0.5^6 + 0.5 \bullet 20 \bullet 0.5^6$$
$$= 0.5$$

意见 7：（对应批注 19）等效的前提是 mj 道两级记分试题是彼此独立，才能连乘，而且作者在编写 EM 算法时，也引用了 Bock 和 Aitkin（1981）提出的假设：项目之间相互独立。但实际中的多级题目，能够满足这个前提要求吗？每个分数等级之间的知识应该是有内在逻辑关系的，从现实角度考虑，很难将一道多级计分题目完全拆成 mj 个相互独立的二级试题。请作者给予解释。否则整个理论体系将会出现问题。

回应：在第一轮回复中，已经详细论述了多级记分题的内在逻辑存在五种类型，以及包括了复杂的知识点关系。但是，目前任何一个多级记分模型都难以完全拟合这些多级记分题的知识点内在逻辑关系。也就是说，如果要具体深究每一个被试在多级记分题的认识水平、知识点关系，而测验中又包含了多个多级记分题，那么目前任何一个 IRT 模型难以进行测量估计，即使在经典测量理论下也是无法区分每一个被试在多级记分题的认识差异。

例如，在某一测验中有一道满分为 10 分的多级记分题，该多级记分试题有 4 个呈递进关系的得分点，还有 6 个相互独立的得分点。同样能力水平（例如得分 85 分）的被试甲、乙在该多级记分试题上得分都为 7 分，那么被试甲、乙在该多级记分题上的认识水平差异，具体是那些评分点掌握了，那些评分点未掌握，在经典测量理论下无法细化评估，在 IRT 理论下也是无法细化评估。

但是人类思维对事物/现象的认识可以采取化繁就简（概括简化）的方式，在其他所有的研究领域是这样，在测量评价研究时也是如此。在多级记分题的测量评分时，测量人员（评分人员）往往是关注被试在各个评分点上的得分是多少，然后累加该多级记分试题的总分；每个评分点的每一分值都是可以累加到多级记分题的试题总分中。因此，在每一个得分点的每一分值对该多级记分题的分值贡献作用，都是相同的，而且都能单独对测验总分（被试能力）产生贡献作用。无论是在经典测量理论下，还是在 IRT 下，对多级记分题的评分过程都是如此。

既然在多级记分题的评分中，在一道试题内部中每一分值的贡献作用是相同的，而且每一分值都能独立对测验总分产生贡献作用。因此，从测验分数的评分作用（测量作用的贡献）的角度看，一道多级计分题目的作用看作成 m 个相互独立的二级计分试题所产生的作用，从理论上是可行的。

意见 8：（对应批注 22）作者前面说到“Logistic 加权模型与其他多级记分模型之间没有共同的前提假设，或者说，没有比较的基础。”

参数体系不一样，确实无法比较。但作者可以换一个思路比较，一个建议：可以拿收集的实证数据，采用拟合指标进行比较，例如-2LL，AIC，BIC。这些指标就是证明新旧模型的有力证据。请作者加上。

本人认为实证数据应该着重描述，模拟研究其实可以弱化一些，毕竟新模型是要指导人们用于实践的，但作者对实证研究部分的阐述过于简单。

回应：以往 IRT 研究中，单、两、三参数 Logistic 模型在实测测验数据中进行模型优劣比较，是因为这几个模型的基本原理、基本的参数体系相同。

而 Logistic 加权模型、GRM 模型是否可以使用实测数据进行模型比较优劣？

作者认为，由于 Logistic 加权模型、GRM 模型的模型假设、模型参数体系不一样，在模拟测验研究中无法比较，同样的，在实测中也同样很难找到同时适合这两个模型假设的实测数据，也无法进行比较。或者说，找到同时适合使用 Logistic 加权模型、GRM 模型的实测数据，在实际中的可能性概率非常小。以下再具体论述。

这里先对实测情境中的多级记分测验进行分类，可以分为两类测验情境：

第一类测验情境是学校中学业成就测验（也包括一些智力测验，能力测验等等最佳行为测验），在这类测验中的多级记分题中几乎都存在以下情况：（1）测验中的选择题评分只有 0 分，X 分（X 大于或等于 2）两种情况；（2）一些多级记分题的部分分数存在得分跨越现象；（3）一些多级记分试题的部分得分上的被试数量太少。如果实测数据出现以上几种情况或其中一种，则在 GRM 模型下是无法进行参数估计, 此时仅适合在 Logistic 加权模型下进行参数估计。在第一类测验情境中，如果在一批测验数据中没有出现以上三种情况，那么从测验数据形式上 Logistic 加权模型、GRM 模型都可以适合，但在实际中找到符合测验数据形式的概率非常小。

第二类测验情境是人格测验中的实测数据，例如 SCL-90 量表、16PF 测验中人格因子部分的测量，在这些测验中所有试题的满分值都相同，比较适合 GRM 模型；而且一些 IRT 经典著作、教材中介绍 GRM 模型时是使用人格测验的例子。但是，这些人格测验的数据不符合 Logistic 加权模型下的基本假设（多级记分题的试题分数起到分数加权作用）。因而，这些人格测验不适合使用 Logistic 加权模型。

因此，前面所论述的第一类情境（绝大多数情况）更适合使用 Logistic 加权模型；第二类情境更适合使用 GRM 模型。因此在第一类情境（绝大多数情况）、第二类情境下，没有必要同时使用两种模型进行比较。

作者对以上第一类测验情境进行再进行以下两种情况假设：

第一种假设：如果对第一类情境下的实测测验进行改造（即在测验的试题命题时就进行改造），使得测验都由多级记分试题组成，而且这些多级记分题的满分值都相同，而且被试得分不出现分数跨越情况，也不出现各个得分上被试数量太少等情况，那么从测验数据形式上 Logistic 加权模型、GRM 模型都可以适合此测验。但是，即使作者设计改造了该测验并收集了这样的实测测验数据，由于该实测测验是经过设计改造了，并没有实测背景、实测测验内容的支持，也“间接”的变成了非真正的实测测验数据，同样会容易受到其他研究者、审稿专家的质疑。

第二种假设：即使在第一类测验情境下找到了这样一个特殊的实测测验情境及实测数据，从测验数据形式上都能够适合 Logistic 加权模型、GRM 模型，并能够计算出 Logistic 加权模型、GRM 模型的-2LL，AIC，BIC 数值大小，但这也只能说明这两个模型在该笔测验数据（该特殊测验情境）上表现的优劣，并不能说明这两个模型在所有多级记分测验（实测测验）中的优劣。因为在实际中，Logistic 加权模型、GRM 模型这两个模型都有各自的模型假设，以及各自适合的测验情境。

总之，作者希望审稿专家在实测数据进行模型对比这一问题上，能保持意见沟通，探讨模型对比的可行性。

对于审稿专家意见“本人认为实证数据应该着重描述，模拟研究其实可以弱化一些，毕竟新模型是要指导人们用于实践的，但作者对实证研究部分的阐述过于简单。”

本文的实证研究部分的内容篇幅虽少，但该实证研究工作其实包含了大量的工作，包括：（1）实测数据整理与收集；（2）在 Logistic 加权模型下，对实测数据的方法分析、Logistic 加权模型的公式推导、实测数据的软件编程测试；（3）在经典测量理论下，实测数据的难度、区分度的计算分析；（4）在 IRT 和 CTT 下的难度参数、区分度参数的对比分析。在原稿中只是为了节省篇幅而论述较为简略。

作者认为，本文已使用 Logistic 加权模型实现了对实测数据的试题参数估计，也就是说，实

测数据的项目参数估计，作为最为关键的应用问题都解决了。因此，其他测验应用问题，包括 Logistic 加权模型下的测验等值、被试能力估计等实证应用问题就比较容易解决。

第三轮

审稿人 1 意见

意见 1：通过认真审阅修改稿，本人认为作者对该模型的思考已经比较全面，基本回答了审稿人提出的问题。收到返修稿的这段时间，本人也在反复思考，前后审阅了多次。首先对一些不清晰的地方提出了问题，然后思索，寻找答案，基本上能够抓住该文章的基本思想和重要信息。正如作者所阐述那样，任何一个模型都不可能尽善尽美，新模型的开发实属不易，作者做了很多工作，我们需要有新的思路存在。

1.修改稿中仍存在一些格式上的问题，请作者认真修改。

2.尽管作者给予了解释，但本人认为利用实证数据进行不同模型之间的比较，还是一项很有必要的工作。

回复：非常感谢审稿专家（包括前两位审稿专家）不辞辛苦提出了许多宝贵的意见，为此研究的完善作出了贡献！同时也感谢审稿专家对此研究给予了肯定评价。

1、对于稿件格式上的问题，作者重新阅读检查全文，进行了检查修改，包括文献引用格式的修改，图形的绘制修改。

2、审稿专家提出“利用实证数据进行不同模型之间的比较”的建议。作者认为，在第二轮回复意见的已有部分内容（详见第二轮回复意见的“审稿专家意见 8”）。正如第二轮回复意见所论述，Logistic 加权模型更适合于第一类测验情境（即学校学业成就测验、智力测验，能力测验等最佳行为测验）；GRM 模型更适合于第二类测验情境（人格测验、态度倾向测验等典型行为测验）。

随着理论研究、实践应用的发展，今后在可能会存在同时符合这两个模型假设交集的实测测验数据，但目前尚未找到这样的实测测验数据。而且，即使在测验实践中存在这样模型假设交集的测验情境，但是，无论这两种模型的比较结果如何，都不影响这两个模型各自的基本假设、应用情境。

附件 1　多级记分题的评分点类型

在心理与教育测验中特别是教育成就测验、智力测验的多级记分试题的评分点类型，可以归纳为以下五种类型：

第一类：单纯的加权形式的多级记分试题（Logistic 加权模型可适合）。例如英语考试中，词汇题目往往是 1 题 1 分，而阅读理解试题往往是 1 题 2 分；高中物理测验中，单选题 1 题 4 分，填空题 1 题 3 分。以上这这种情况都是一种单纯的试题分数加权。

第二类：并列形式的多级记分试题（Logistic 加权模型可适合）。在有些多级记分试题中，各个得分小点之间没有相互依赖关系，也没有递进关系，此时的多级记分试题可以“拆分为”，或者视作为多个 1，0 记分试题，可以直接使用 1，0 记分模型来适合。而此时，使用多级记分模型也是可以的，因为多级记分模型的特例，即是 1，0 记分模型，也就是说包含了 1，0 记分模型。例如：一道简答题，满分为 5 分，包含五个小点，答对任意一个小点都给 1 分，答对其中两个小点就给 2 分，依此类推。

第三类：存在一定相依关系的多级记分试题（Logistic 加权模型可勉强适合，但可能损失一小部分的测量精度）。在语文学科、或英语的阅读理解试题中，一篇阅读短文并据此附有 5 道

小题，此 5 道小题的作答之间，可能是存在完全的并列关系，也可能存在一定的相依关系。相依关系，不同于并列关系形式，也不同于递进关系形式。此时，可以使用 1，0 记分试题下的相依反应模型。但是在许多研究中发现，这样相依关系往往比较弱，往往可以忽略这种相依关系，而直接视作为并列关系的多级记分试题，这种情况下可以使用多级记分模型来处理（Sereci，Wainer，& Thissen，1991）。

第四类：递进形式的多级记分试题（Logistic 加权模型可适合）。在有些多级记分试题中，各个得分小点之间存在着先后递进的依赖关系，而且是必需先作答完成前面一个得分点，才能进行下一个得分点的作答。（这里的得分点可以是多级记分的一个小题，亦可以是多级记分试题中一个小题中还包含了多个得分点。）例如，一道试题分为 3 个递进关系的得分点，作答完成第 1 个得分点时可以得 1 分，作答完成第二个得分点时，可再得 1 分（累计得 2 分），作答完成第三得分点时，可以再得 1 分（累计得 3 分）。在这种情况下，Samejima 等级反应模型、评定量表模型、分部评分模型都可以适合。但是，在实际考试中，命题人往往会根据知识重要性和所需思考过程的复杂程度给予试题得分点赋予得分权重，例如，作答完成第 1 个得分点时可以得 2 分，作答完成第二个得分点时，可再得 4 分（累计得 6 分），作答完成第三得分点时，可以再得 3 分（累计得 9 分）。此时，被试得分只有 4 种得分的可能，即 0 分，2 分，6 分，9 分。在这种情况下，得分出现断层情况，而且需要表现出得分权重时，以往的多级记分模型都不能适合这种情况。

第五类：以上四种基本类型中的两种或两种以上类型混合而成的多级记分评分点类型。

在实际测验中，一份测验可能会同时包含这五种类型，或者其中的几种类型。

附件 2　Logistic 加权模型的项目参数估计 EM 算法推导

Logistic 加权模型的项目特征函数描述，与等级反应模型等一些多级记分模型一样的，都有两种方式来表示 Logistic 加权模型的项目特征曲线函数：

第一种方式：被试恰好得$u_{\alpha j}$分的项目特征函数为$C_{m_j}^{u_{\alpha j}} P_{\alpha j}{}^{u_{\alpha j}} Q_{\alpha j}{}^{m_j-u_{\alpha j}}$，分别为：

$$C_{m_j}^{0} Q_{\alpha j}{}^{m_j}，\ C_{m_j}^{1} P_{\alpha j}{}^{1} Q_{\alpha j}{}^{m_j-1}，\ C_{m_j}^{2} P_{\alpha j}{}^{2} Q_{\alpha j}{}^{m_j-2}，\ \ldots，\ C_{m_j}^{k} P_{\alpha j}{}^{k} Q_{\alpha j}{}^{m_j-k} \ldots，\ C_{m_j}^{m_j} P_{\alpha j}{}^{m_j}； \qquad (1)$$

第二种方式：被试得$u_{\alpha j}$（$0 \le u_{\alpha j} \le m_j$）分或$u_{\alpha j}$分以上的项目特征函数为：

$$\sum_{u_{\alpha j}=m_j}^{m_j} C_{m_j}^{u_{\alpha j}} P_{\alpha j}{}^{u_{\alpha j}} Q_{\alpha j}{}^{m_j-u_{\alpha j}} \qquad (2)$$

其中：

$$P = 1/[1+\exp(-1.7a(\theta-b))]， \qquad (3)$$

$$Q = 1-P \qquad (4)$$

Logistic 加权模型可以使用 MMLE/EM 算法估算出两级记分试题的区分度参数、难度参数，和多级记分试题的区分度参数、平均难度参数，其基本过程与两级记分试题的参数估计过程相似。以下简要论述 Logistic 加权模型下 MMLE/EM 算法的基本过程。

在两级和多级记分试题组成的测验中，根据全体被试在所有项目上的作答数据矩阵U，建立被试作答的边际似然函数：

$$L = \prod_{\alpha=1}^{N} \int \prod_{j=1}^{M} P_{ij}{}^{u_{\alpha j}} Q^{m_j-u_{\alpha j}} g(\theta) d\theta \qquad (5)$$

其中m_j为试题j的满分，且$m_j \ge 1$，当$m_j = 1$为两级记分试题；$u_{\alpha j}$为第α被试在试题j的

得分，且$0 \le u_{\alpha j} \le m_j$。在函数（5）两边求对数，并分别求$a_j$，$b_j$的导数，并令$a_j$，$b_j$的导数等于 0，得到$a_j$，$b_j$的方程组。这里的$b_j$两级记分试题的难度参数，或多级记分试题的平均难度参数。

$$\frac{\partial \ln L}{\partial a_j} = 0 \qquad (6)$$

$$\frac{\partial \ln L}{\partial b_j} = 0 \qquad (7)$$

依据 Bock & Aitkin（1981）提出的假设：项目之间相互独立，被试之间相互独立，项目与被试相互独立。因此，可以逐个项目迭代求解项目参数。以上方程（6）式、（7）式里出现积分是采用数值积分的方式。设定$g(\theta)$为正态分布函数，对被试能力求积分，把正态分布分为q个结点$x_1, x_2 \ldots x_k \ldots x_q$，求出对应的结点系数$A(x_1),\ A(x_1), \cdots, A(x_k)\ \cdots, A(x_q)$。把这些边际似然方程（6）式、（7）式,化为数值积分形式，分别为：

$$\frac{\partial \ln L}{\partial a_j} = \sum_{k=1}^{q}\sum_{\alpha=1}^{N} D \cdot (u_{\alpha j} - m_j \cdot P_j(x_k)) \cdot (x_k - b_j) \cdot h(x_k \mid u_\alpha) = 0 \qquad (8)$$

$$\frac{\partial \ln L}{\partial b_j} = \sum_{k=1}^{q}\sum_{\alpha=1}^{N} -D \cdot a_j \cdot (u_{\alpha j} - m_j \cdot P_j(x_k)) \cdot h(x_k \mid u_\alpha) = 0 \qquad (9)$$

其中 $h(x_k \mid u_\alpha) = \dfrac{\prod_{j=1}^{M} P_j(x_k)^{u_{\alpha j}} \cdot Q_j(x_k)^{m_j - u_{\alpha j}} \cdot A_j(x_k)}{\sum_{q=1}^{k}\prod_{j=1}^{M} P_j(x_k)^{u_{\alpha j}} \cdot Q_j(x_k)^{m_j - u_{\alpha j}} \cdot A_j(x_k)}$

且令 $L(x_k) = \prod_{j=1}^{M} P_j(x_k) Q_j(x_k)$，则 $h(x_k \mid u_\alpha) = \dfrac{L(x_k) \cdot A_j(x_k)}{\sum_{q=1}^{k} L(x_k) \cdot A_j(x_k)}$

根据 EM 算法，进一步建立“人工数据”：

$$\overline{f_k} = \sum_{\alpha=1}^{N} h(x_k \mid u_\alpha) = \sum_{i=1}^{N} \frac{L(x_k) \cdot A_j(x_k)}{\sum_{q=1}^{k} L(x_k) \cdot A_j(x_k)} \qquad (10)$$

$$\overline{r_{jk}} = \sum_{\alpha=1}^{N} \frac{u_{\alpha j}}{m_j} \cdot h(x_k \mid u_\alpha) = \sum_{i=1}^{N} \frac{\frac{u_{\alpha j}}{m_j} \cdot L(x_k) \cdot A_j(x_k)}{\sum_{q=1}^{k} L(x_k) \cdot A_j(x_k)} \qquad (11)$$

其中 $x_1, x_2 \ldots x_k \ldots x_q$ 为正态分布的求积结点，而 $A(x_1),\ A(x_1), \cdots, A(x_k)\ \cdots, A(x_q)$ 为求积系数。在项目参数估计迭代求解过程中逐步

调整的，计算公式为 $A_j(x_k)=\overline{f_k}/N$ 。由于在 $\overline{f_k}$ 中没有项目足码 j，即 $\overline{f_k}$ 与项目的 j 无关，表示容量为 N 的总体中期望能力为 x_k 的被试的人数；而 $\overline{r_{jk}}$ 与项目 j 有关，表示该总体中具有能力为 x_k 的被试答对第 j 项目的人数。用 $\overline{f_k}$ 和 $\overline{r_{jk}}$ 改写（8）式、（9）式方程组可得：

$$f_1=\frac{\partial \ln L}{\partial a_j}=\sum_{k=1}^{q}D\cdot(\overline{r_{jk}}-\overline{f_k}\cdot P_j(x_k))\cdot(x_k-b_j)=0 \qquad (12)$$

$$f_2=\frac{\partial \ln L}{\partial b_j}=\sum_{k=1}^{q}-D\cdot a_j\cdot(\overline{r_{jk}}-\overline{f_k}\cdot P_j(x_k))=0 \qquad (13)$$

要解（12）式、（13）式方程组，需要由 f_1， f_2 方程组对 a_j， b_j 分别求一阶导数、一阶偏导，并得到 2*2 的矩阵，求偏导的公式可参考 IRT 文献(Baker, 2004; 漆书青, 戴海崎, & 丁树良, 2002) ，或者参考求偏导的高数教材。在（12）式、（13）式是非线性方程，使用 Newton-Raphson 迭代方法需要预先计算初值。项目区分度参数的初值为 $a_j=r_{b_j}/(\sqrt{1-r_{b_j}^{\ 2}}\cdot\sqrt{m_j})$ ，难度参数的初值为 $b_j=z_j/r_{b_j}$ ，其中 $r_{b_j}=r_{pbj}\sqrt{p_j\cdot(1-p)_j}/y(z_j)$ ， z_j 为项目得分率 p 转换后标准正态分数， r_{pbj} 是项目得分与试卷总分的点二列相关系数， $y(z_j)$ 是正态分布密度函数值。

在（12）式、（13）式中， $\overline{f_k}$ 和 $\overline{r_{jk}}$ 依赖于 $h(x_k\mid u_\alpha)$，而 $h(x_k\mid u_\alpha)$ 又含有 $P_{j(x_k)}$， $P_{j(x_k)}$ 中含有参数 a_j，b_j，所以 $\overline{f_k}$ 和 $\overline{r_{jk}}$ 依赖于项目参数 a_j，b_j。经过 Newton-Raphson 方法迭代解出 a_j，b_j 的值，解出的 a_j， b_j 的值可以计算新一轮迭代的 $\overline{f_k}$ 和 $\overline{r_{jk}}$ 。在 Newton-Raphson 方法迭代中，$\overline{f_k}$ 和 $\overline{r_{jk}}$ 与项目参数估计值相互迭代估计，逐步估计出项目参数值 a_j， b_j 。

附件 3：GRM 下多级记分题的项目信息量的计算

在 GRM 下第 j 题对被试 α 的项目信息量为：

$$I_j(\theta_\alpha)=a_j^{\ 2}\sum_{t=0}^{f_j}\left[P_{\alpha j,t}^{\ *}(\theta_\alpha)-P_{\alpha j,t+1}^{\ *}(\theta_\alpha)\right]\left[1-P_{\alpha j,t}^{\ *}(\theta_\alpha)-P_{\alpha j,t+1}^{\ *}(\theta_\alpha)\right]^2$$

$P_{\alpha j,t}^{\ *}(\theta_\alpha)$ 为第 α 个被试在第 j 题上得 t 分及 t 分以上的概率， f_j 为第 j 题的满分值， a_j 为第 j 题的区分度。

这里在 GRM 模型下设计有 8 道试题，这些试题区分度 a 都为 1.7（此区分度值包含常量 D＝1.7），难度参数分别为：

第 1 题 1 个难度参数，难度参数为 0.5;

第 2 题，3 个难度参数，分别为，-0.5，0.5，1.5；难度中值为 0.5。

第 3 题，4 个难度参数，4 个难度参数分别为，-0.5，0.0，1.0，1.5；难度中值为 0.5。

第 4 题，5 个难度参数，五个难度参数分别为，-1.5，-0.5， 0.5，1.5，2.5; 难度中值为 0.5。

第 5 题，7 个难度参数分别为，-2.5，-1.5，-0.5， 0.5，1.5，2.5，3.5;

第 6 题，9 个难度参数分别为，-2.5，-1.5，-0.5， 0.5，1.5，2.5，3.5，4.5，5.5;

第 7 题，11 个难度参数分别为，-2.5，-1.5，-0.5， 0.5，1.5，2.5，3.5，4.5，5.5，6.5，7.5;

第 8 题，17 个难度参数分别为，-2.5，-2.0，-1.5，-1.0，-0.5， 0.0，0.5，1.0，1.5，2.0，2.5，3.0，3.5，4.0，4.5，5.0，5.5;

第 9 题，18 个难度参数分别为，-2.5，-1.5，-0.5，0.5，1.5，2.5，3.5，4.5， 5.5，5.8，6.1，6.4， 6.7，7.0，7.3，7.6，7.9，8.2。

计算以上 9 道试题的项目信息量，可得图 1：

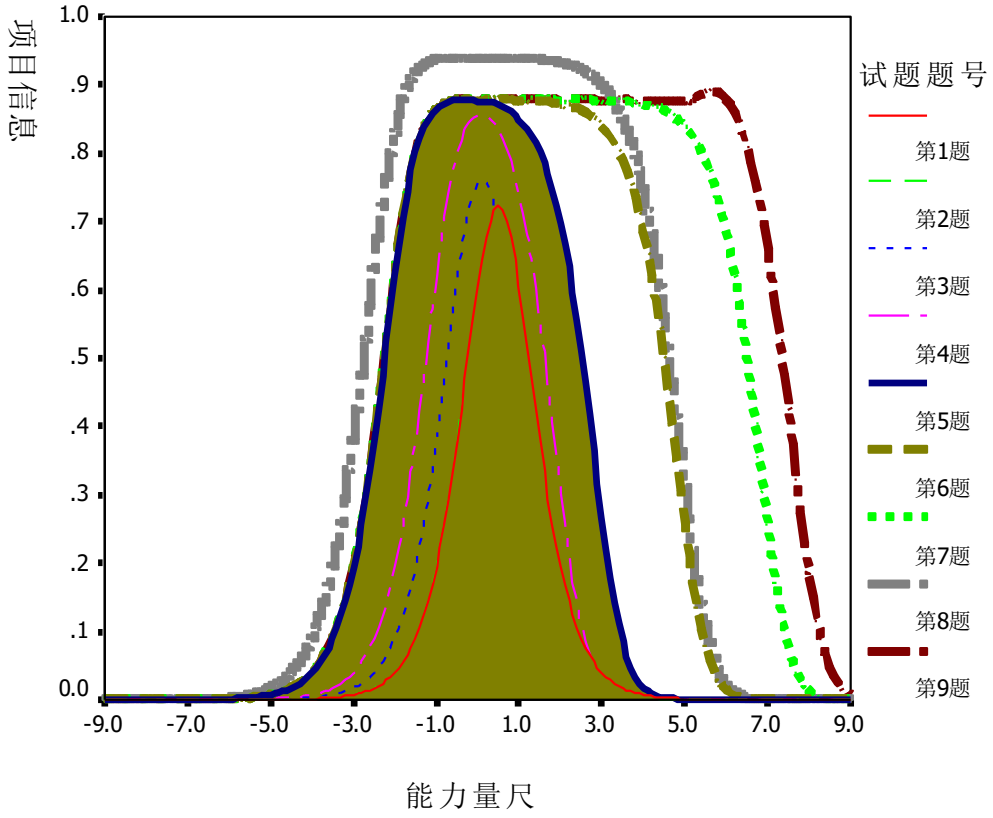

图1 九道试题在GRM下的项目信息量

（1）以第 1 题为基础，对比第 2、3、4、5 题的难度参数，这些试题难度均值都在 0.5 左右，而且都是围绕着 0.5 增加试题分数等级。由图可知，随着试题满分值的增加，试题的项目信息量增加并不是很多，第 1 题的项目信息量最大值为 0.7225，而第 5 题的项目信息量的最大值为 0.8785，也就是说在 GRM 下多个记分等级，并不能有效的增加试题项目信息量。而且第 2 题为 3 个难度参数等级，但其项目信息量比第 1 题的项目信息量还要少。

以上这些说明，如果在一个多级记分试题的两端增加多级记分等级，虽然能增加少量的项目信息量，但增加的幅度相对较小。对比第 3 题 4 个难度参数，第 5 题 7 个难度参数，难度参数等级几乎增加了一倍，但项目信息量的最大值只增加了 20%左右。

（2）对比第 5、6、7 题，第 6、7 题是在第 5 题的基础上，在高难度一端增加试题难度等级。由图可知，第 6、7 题在能力量尺上的项目信息量的最大值基本没有改变，不同的是第 6、7 题的顶部变宽了。以上说明，如果在一个多级记分试题的高难度一端增加多级记分等级，虽然项目信息量的最大值几乎没有增加，但在能力量尺上的最大信息量顶部宽度变宽了。

（3）比较第 6 题和第 8 题，在 9 个记分等级的基础上，在第 6 题的多个难度等级之间插入式的增加 8 个记分等级，发现项目信息量最大值由第 6 题的 0.8785 增大到第 8 题的 0.9406，但增大幅度仅为 8%。同时项目信息量曲线的顶部稍微加宽了。

（4）比较第 6 题和第 9 题，在 9 个记分等级的基础上，在第 6 题的多个难度等级的右侧高难度的一端增加 8 个记分等级，发现项目信息量最大值的增加幅度几乎没有变化，在能力量尺上的最大信息量顶部宽度变宽了。当然，在能力量尺区间[5.5,7.5]之间的顶部最大值稍微增大了 2%，这是因为第 9 题在[5.5,8.5]之间密集的增加了多个难度参数导致的。

总之，从项目信息量的角度来看，一道多级记分试题提供的项目信息量仅仅比一道难度相当的两级记分试题的项目信息量增加小部分，而且得分等级增加 1 倍，而项目信息量最大值的增加幅度不超过 20%。而且如果在难度量尺的高难度一端增加分数等级，项目信息量的最大值几乎不变，而项目信息量的最大值的顶部变宽。